Time Series Outlier Detection

Tingyi Zhu

July 28, 2016
Outline

- Time Series Basics
- Outliers Detection in Single Time Series
- Outlier Series Detection from Multiple Time Series
- Demos
Time Series Basics
First-order Autoregression

A model denoted as AR(1), in which the value of $X$ at time $t$ is a linear function of the value of $X$ at time $t - 1$:

$$X_t = \phi X_{t-1} + \varepsilon_t$$ \hspace{1cm} (1)

Assumptions:

- $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma)$, stochastic term.
- $\varepsilon_t$ is independent of $X_t$. 
General Autoregressive Model

AR(p):

\[ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \]

\[ = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t \]

\[ = \sum_{i=1}^{p} \phi_i B^i X_t + \varepsilon_t \]

where we use the backshift operator \( B \) (\( BX_t = X_{t-1}, B^k X_t = X_{t-k} \)).

Alternative notation:

\[ \phi(B)X_t = \varepsilon_t \]

\( \phi(B) \) is a polynomial of \( B \),

\[ \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 1 - \sum_{i=1}^{p} \phi_i B^i \]
Moving Average

- Another approach for modeling univariate time series
- $X_t$ depends linearly on its own current and previous stochastic terms
- MA(1):
  
  $$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1}$$

- MA(q):
  
  $$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}$$
\( \theta_1, \ldots, \theta_q \): parameters of MA model

\( \varepsilon_t, \ldots, \varepsilon_{t-q} \): stochastic terms

Using backshift operator \( B \), model simplified as

\[
X_t = (1 + \theta_1 B + \cdots + \theta_q B^q) \varepsilon_t \\
= (1 + \sum_{i=1}^{q} \theta_i B^i) \varepsilon_t \\
= \theta(B) \varepsilon_t
\]
ARMA Model

- A model consists of both autoregressive (AR) part and moving average (MA) part:

\[ X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \]  

(2)

referred to as the ARMA(p,q) model.

- \( p \): the order of the autoregressive part
- \( q \): the order of the moving average part

- More concisely, using backshift operator \( B \), (2) becomes:

\[ \phi(B)X_t = \theta(B)\varepsilon_t \]
In short, a time series is stationary if its statistical properties are all constant over time.

To mention some properties:

- Mean: \( E[X_t] = E[X_s] \) for any \( t, s \in \mathbb{Z} \),
- Variance: \( \text{Var}[X_t] = \text{Var}[X_s] \) for any \( t, s \in \mathbb{Z} \),
- Joint distribution:
  \[
  \text{Cov}(X_t, X_{t+1}) = \text{Cov}(X_s, X_{s+1}) \text{ for any } t, s \in \mathbb{Z}.
  \]
Stationary Time Series

Non-stationary Time Series
Requirements for a Stationary Time Series

- **AR(1)** \( X_t = \phi X_{t-1} + \varepsilon_t: \quad |\phi| < 1 \)

- **AR(p)** \( \phi(B)X_t = \varepsilon_t: \)
  
  All the roots of \( \phi(z) = 0 \) are outside unit circle.

- **MA models are always stationary**

- **ARMA(p,q)** \( \phi(B)X_t = \theta(B)\varepsilon_t: \)
  
  All the roots of \( \phi(z) = 0 \) are outside unit circle.
Non-stationary time series

- Trend effect
- Seasonal effect

Figure: Monthly totals of international airline passengers, 1949 to 1960.
Think of a more general time series formulation including both trend and seasonal effect:

\[ X_t = T_t + S_t + E_t \]  (3)

- \( X_t \) is data point at time \( t \)
- \( T_t \) is the trend component at time \( t \)
- \( S_t \) is the seasonal component at time \( t \)
- \( E_t \) is the remainder component at time \( t \) (containing AR and MA terms)
Series with Trend, examples:

- Assuming no seasonal effect, i.e. $S_t = 0$

  - Linear trend:
    
    $$X_t = 2t + 0.5X_{t-1} + \varepsilon_t$$

  - Quadratic trend:
    
    $$X_t = 2t + t^2 + 0.5X_{t-1} + \varepsilon_t$$

- Goal: remove the trend, to transform the series to be stationary

- Solution: lag-1 differencing
Differencing and Trend

Define the lag-1 difference operator,

\[ \nabla X_t = X_t - X_{t-1} = (1 - B)X_t, \]

where \( B \) is the backshift operator.

- If \( X_t = \beta_0 + \beta_1 t + E_t \), then
  \[ \nabla X_t = \beta_1 + \nabla E_t. \]

- If \( X_t = \sum_{i=0}^{k} \beta_i t^i + E_t \), then
  \[ \nabla^k X_t = (1 - B)^k X_t = k! \beta_k + \nabla^k E_t. \]

We call \( \nabla^k \) kth lag-1 difference operator.
Lag-1 Differencing

S&P 500 Quote Year-To-Date

S&P 500 YTD Lag-1 Differencing
Series with Seasonal Effect, example:

- For quarterly data, with possible seasonal (quarterly) effects, we can define indicator function $S_j$. For $j = 1, 2, 3, 4$,

$$S_j = \begin{cases} 
1 & \text{if observation is in quarter } j \text{ of a year,} \\
0 & \text{otherwise.}
\end{cases}$$

- A model with seasonal effects could be written as

$$X_t = \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 S_4 + \varepsilon_t$$

- Goal: remove the seasonal effects

- Solution: lag-$s$ differencing, where $s$ is the number of seasons
Differencing and Seasonal Effects

Define the lag-$s$ difference operator,

$$\nabla_s X_t = X_t - X_{t-s} = (1 - B^s)X_t,$$

where $B$ is the backshift operator.

If $X_t = T_t + S_t + E_t$, and $S_t$ has period $s$ (i.e. $S_t = S_{t-s}$ for all $t$), then

$$\nabla_s X_t = (1 - B^s)X_t = T_t - T_{t-s} + \nabla_s E_t.$$
Non-seasonal ARIMA

- $S_t = 0$

- ARIMA stands for Auto-Regressive Integrated Moving Average, ARMA integrated with differencing.

- A nonseasonal ARIMA model is classified as ARIMA($p,d,q$), where
  - $p$ is the order of AR terms,
  - $d$ is the number of nonseasonal differences needed for stationarity,
  - $q$ is the order of MA terms.
Non-seasonal ARIMA, Cont.

- Recall ARMA(p,q):
  \[ \phi(B)X_t = \theta(B)\varepsilon_t, \]
  - \( \phi(B) \) and \( \theta(B) \) are polynomials of \( B \) of order \( p \) and \( q \).
  - Stationary requirement: all roots of \( \phi(z) = 0 \) outside unit circle.

- ARIMA(p,d,q):
  \[ \phi(B)(1 - B)^dX_t = \theta(B)\varepsilon_t, \]
  - \( X_t \) is not stationary. Why?
  - \( Z_t = (1 - B)^dX_t \) is ARMA(p,q), is stationary.
A seasonal ARIMA model is classified as

\[ ARIMA(p, d, q) \times (P, D, Q)_m \]

- \( p \) is the order of AR terms,
- \( d \) is the number of nonseasonal differences,
- \( q \) is the order of MA terms.
- \( P \) is the order of seasonal AR terms,
- \( D \) is the number of seasonal differences,
- \( Q \) is the order of seasonal MA terms.
- \( m \) is the number of seasons.
Example: $ARIMA(1, 1, 1) \times (1, 1, 1)_4$
General ARIMA

- The ARIMA model can be generalized as follow:

\[ \phi(B)\alpha(B)X_t = \theta(B)\varepsilon_t, \]

- \( \phi(B) \): autoregressive polynomial, all roots outside unit circle
- \( \alpha(B) \): differencing filter renders the data stationary, all roots on the unit circle
- \( \theta(B) \): moving average polynomial, all roots outside unit circle (to assure \( \theta(B) \) is invertible.

- Alternatively,

\[ X_t = \frac{\theta(B)}{\phi(B)\alpha(B)}\varepsilon_t. \]
Outliers Detection in Single Time Series
Automatic Detection Procedure

- Based on the framework of ARIMA models
- R package tsoutlier written by YAHOO in 2014
Types of Outliers

- General representation: $L(B)I_t(t_j)$
  - $L(B)$: a polynomial of lag operator $B$
  - $I_t(t_j) = 1$ there’s outlier at time $t = t_j$, and 0 otherwise.

- Types of outliers:
  - Additive Outliers (AO): $L(B) = 1$;
  - Level Shift (LS): $L(B) = \frac{1}{1-B}$;
  - Temporary Change (TC): $L(B) = \frac{1}{1-\delta B}$;
  - Seasonal Level Shift (SLS): $L(B) = \frac{1}{1-B^s}$;
  - Innovational Outliers (IO): $L(B) = \frac{\theta(B)}{\phi(B)\alpha(B)}$. 
Types of Outliers

AO: additive outlier

TC: temporary change

LS: level shift

SLS: seasonal level shift (quarterly data)

IO: innovational outlier. ARMA(3,2)

IO: innovational outlier. ARIMA(0,1,1)(0,1,1)
ARIMA model:

\[ X_t = \frac{\theta(B)}{\phi(B)\alpha(B)} \varepsilon_t. \]

Model with outliers at time \( t_1, t_2, \ldots, t_m \):

\[ X_t^* = \sum_{j=1}^{m} \omega_j L_j(B) I_t(t_j) + \frac{\theta(B)}{\phi(B)\alpha(B)} \varepsilon_t. \]

- \( L_j(B) \) depends on pattern of the \( j \)th outlier
- \( I_t(t_j) = 1 \) there's outlier at time \( t = t_j \), and 0 otherwise.
- \( \omega_j \) denotes the magnitude of the \( j \)th outlier effect
Effect of One Outlier

- Assume the time series parameters are known, we examine the effect of one outlier:

\[ X_t^* = \omega L(B)l_t(t_1) + \frac{\theta(B)}{\phi(B)\alpha(B)}\varepsilon_t \]

- Define polynomial \( \pi(B) \) as:

\[ \pi(B) = \frac{\phi(B)\alpha(B)}{\theta(B)} = 1 - \pi_1B - \pi_2B - \cdots, \]

- Contaminated by the outlier, the estimated residual \( \hat{e}_t \) becomes

\[ \hat{e}_t = \pi(B)X_t^* \]

(Without outlier, \( \hat{e}_t = \pi(B)X_t \).)
For the four types of outliers,

- **IO**: $\hat{e}_t = \omega l_t(t_1) + \varepsilon_t$
- **AO**: $\hat{e}_t = \omega \pi(B) l_t(t_1) + \varepsilon_t$
- **LS**: $\hat{e}_t = \omega \frac{\pi(B)}{1-B} l_t(t_1) + \varepsilon_t$
- **TC**: $\hat{e}_t = \omega \frac{\pi(B)}{1-\delta B} l_t(t_1) + \varepsilon_t$

Alternatively,

$$\hat{e}_t = \omega x_{i,t} + \varepsilon_t, \quad t = t_1, t_1 + 1, \ldots \text{ and } i = 1, 2, 3, 4$$

- $x_{i,t} = 0$ for all $i$ and $t < t_1$,
- $x_{i,t} = 1$ for all $i$,
- $x_{1,t_1+k} = 0$, $x_{2,t_1+k} = -\pi_k$,
- $x_{3,t_1+k} = 1 - \sum_{j=1}^{k} \pi_j$, $x_{4,t_1+k} = \delta^k - \sum_{j=1}^{k-1} \delta^{k-j} \pi_j - \pi_k$.

A simple linear regression!
Estimate of $\omega$

The least square estimate doe the effect of a single outlier at $t = t_1$ can be expressed as

$$\hat{\omega}_{IO}(t_1) = \hat{e}_{t_1}$$

$$\hat{\omega}_{AO}(t_1) = \frac{\sum_{i=t_1}^{n} \hat{e}_i x_{2t}}{\sum_{i=t_1}^{n} x_{2t}^2}$$

$$\hat{\omega}_{LS}(t_1) = \frac{\sum_{i=t_1}^{n} \hat{e}_i x_{3t}}{\sum_{i=t_1}^{n} x_{3t}^2}$$

$$\hat{\omega}_{TC}(t_1) = \frac{\sum_{i=t_1}^{n} \hat{e}_i x_{4t}}{\sum_{i=t_1}^{n} x_{4t}^2}. $$
From regression analysis, we have

\[
\frac{\hat{\omega} - \omega}{\hat{\sigma}_a} \left( \sum_{t=t_1}^{n} x_{i,t}^2 \right)^{1/2} \sim N(0, 1),
\]

where \( \hat{\sigma}_a \) is the estimation of residual standard deviation.

We want to test whether \( \omega = 0 \), then the following statistics are approximately \( N(0, 1) \):

\[
\hat{\tau}_{IO}(t_1) = \frac{\hat{\omega}_{IO}(t_1)}{\hat{\sigma}_a}
\]

\[
\hat{\tau}_{AO}(t_1) = \left\{ \frac{\hat{\omega}_{AO}(t_1)}{\hat{\sigma}_a} \right\} \left( \sum_{t=t_1}^{n} x_{2t}^2 \right)^{1/2}
\]

\[
\hat{\tau}_{LS}(t_1) = \left\{ \frac{\hat{\omega}_{LS}(t_1)}{\hat{\sigma}_a} \right\} \left( \sum_{t=t_1}^{n} x_{3t}^2 \right)^{1/2}
\]

\[
\hat{\tau}_{TC}(t_1) = \left\{ \frac{\hat{\omega}_{TC}(t_1)}{\hat{\sigma}_a} \right\} \left( \sum_{t=t_1}^{n} x_{4t}^2 \right)^{1/2}.
\]
Procedure in the Presence of Multiple Outliers

In the presence of multiple outliers, recall the model

\[ X_t^* = \sum_{j=1}^{m} \omega_j L_j(B) l_t(t_j) + \frac{\theta(B)}{\phi(B) \alpha(B)} \varepsilon_t. \]

where \( \hat{\sigma}_a \) is the estimation of residual standard deviation.

The estimated residual becomes

\[ \hat{e}_t = \sum_{j=1}^{m} \omega_j \pi(B)L_j(B) l_t(t_j) + \varepsilon_t \]
Stage 1: Joint Estimation of Outlier Effect and Model Parameters

- Fitting the series by an ARIMA model (forecast package in R), obtain initial parameter \((\phi(B), \theta(B), \alpha(B))\) estimation of the model.

- Detect outliers one by one sequentially

I.2. For \(t = 1, \ldots, n\), compute \(\hat{\tau}_{\text{IO}}(t), \hat{\tau}_{\text{AO}}(t), \hat{\tau}_{\text{LS}}(t), \) and \(\hat{\tau}_{\text{TC}}(t)\) in (14) using the residuals obtained from I.1, and let \(\eta_t = \max\{|\hat{\tau}_{\text{IO}}(t)|, |\hat{\tau}_{\text{AO}}(t)|, |\hat{\tau}_{\text{LS}}(t)|, |\hat{\tau}_{\text{TC}}(t)|\}\). If \(\max_t \eta_t = |\hat{\tau}_{\text{tp}}(t_1)| > C\), where \(C\) is pre-determined critical value, then there is a possibility of a type \(tp\) outlier at \(t_1\); \(tp\) may be IO, AO, LS, or TC.
Stage 2: Initial Parameter Estimation and Outlier Detection

II.1. Suppose that \( m \) time points \( t_1, t_2, \ldots, t_m \) are identified as possible outliers of various types. The outlier effects \( \omega_j \)'s can be estimated jointly using the multiple regression model described in (20), where \( \{\hat{\epsilon}_t\} \) is regarded as the output variable and \( \{L_j(B)I_t(t_j)\} \) are the input variables.

II.2. Compute the \( \hat{r} \) statistics of the estimated \( \omega_j \)'s, where \( \hat{r}_j = \hat{\omega}_j / \text{std}(\hat{\omega}_j), j = 1, \ldots, m \). If \( \min_j|\hat{r}_j| = \hat{r}_\nu \leq C \), where \( C \) is the same critical value used in step I.2, then delete the outlier at time point \( t_\nu \) from the set of the identified outliers and go to step II.1 with the remaining \( m - 1 \) outliers. Otherwise, go to step II.3.
II.3. Obtain the adjusted series by removing the outlier effects, using the most recent estimates of $\omega_j$'s at step II.1. In other words, remove only the outlier effects that are significant based on the iterations of steps II.1 and II.2.

II.4. Compute the maximum likelihood estimates of the model parameters based on the adjusted series obtained at step II.3. If the relative change of the residual standard error from the previous estimate is greater than $\epsilon$, go to step II.1 for further iterations;
Outlier Series Detection from Multiple Time Series
Detect Anomalous Series

- **Goal:** efficiently find the least similar time series in a large set
- **Motivation:** Internet companies monitoring the servers (CPU, Memory), find unusual behaviors
Detect Anomalous Series


- Approach: Extract features from time series, PCA

- R package anomalous

- Test on real data from YAHOO email server, 80% accuracy compared to 40% from previous methods
Step 1: Extract Features from Time Series

- 15 features selected, each captures the global information of time series

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Mean.</td>
</tr>
<tr>
<td>Var</td>
<td>Variance.</td>
</tr>
<tr>
<td>ACF1</td>
<td>First order of autocorrelation.</td>
</tr>
<tr>
<td>Trend</td>
<td>Strength of trend.</td>
</tr>
<tr>
<td>Linearity</td>
<td>Strength of linearity.</td>
</tr>
<tr>
<td>Curvature</td>
<td>Strength of curvature</td>
</tr>
<tr>
<td>Season</td>
<td>Strength of seasonality.</td>
</tr>
<tr>
<td>Peak</td>
<td>Strength of peaks.</td>
</tr>
<tr>
<td>Trough</td>
<td>Strength of trough.</td>
</tr>
<tr>
<td>Entropy</td>
<td>Spectral entropy.</td>
</tr>
<tr>
<td>Lumpiness</td>
<td>Changing variance in remainder.</td>
</tr>
<tr>
<td>Spikiness</td>
<td>Strength of spikiness</td>
</tr>
<tr>
<td>Lshift</td>
<td>Level shift using rolling window.</td>
</tr>
<tr>
<td>Vchange</td>
<td>Variance change.</td>
</tr>
<tr>
<td>Fspots</td>
<td>Flat spots using discretization.</td>
</tr>
<tr>
<td>Cpoints</td>
<td>The number of crossing points.</td>
</tr>
<tr>
<td>KLscore</td>
<td>Kullback-Leibler score.</td>
</tr>
<tr>
<td>Change.idx</td>
<td>Index of the maximum KL score.</td>
</tr>
</tbody>
</table>

Table 1: Summary of features used for detecting unusual time series.
Step 2: PCA to reduce dimension

- dim＝15 initially, correlation existing between features
- The first 2 PCs are sufficient, capturing most of the variance

Step 3: Implement multi-dimensional outlier detection algorithm to find outlier series

- Density based
- $\alpha$-hull
Demo